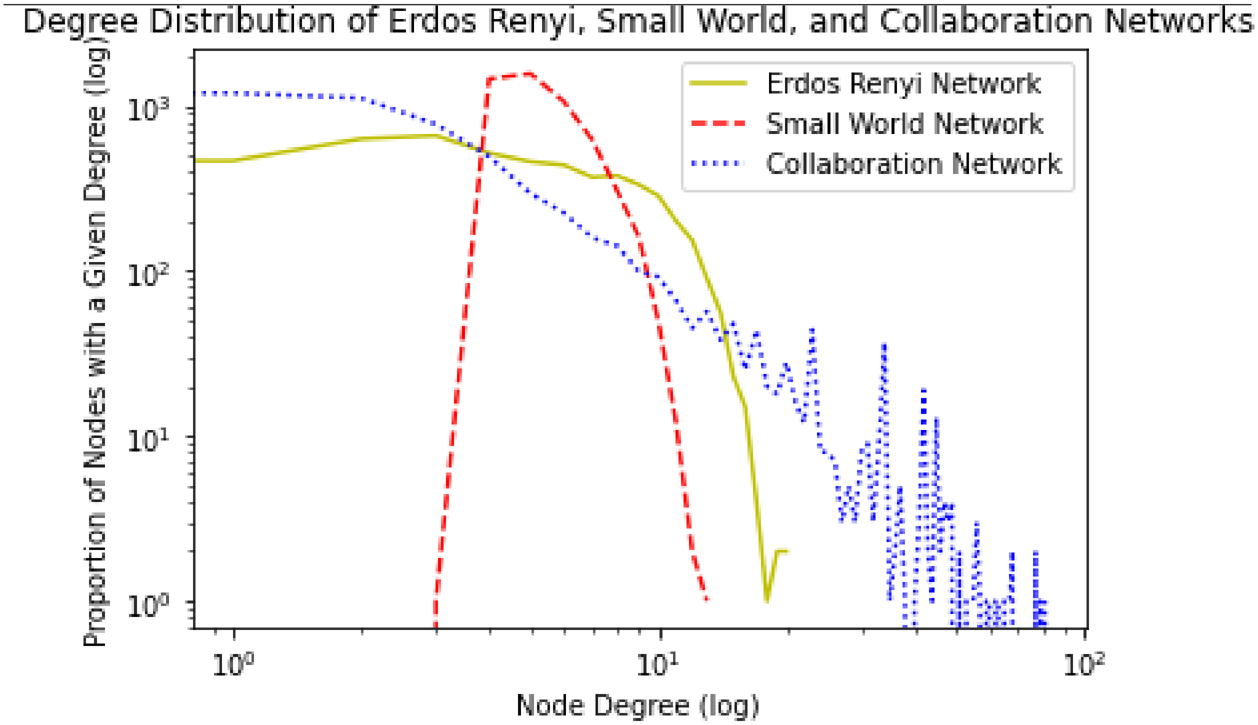


The collaboration graph has a curve that is more stretched out to the right side compared to the random graphs, suggesting it has more nodes with higher degree counts than the random graphs.

Clustering Coefficient for Erdos Renyi Network: 0.001555
Clustering Coefficient for Small World Network: 0.296097
Clustering Coefficient for Collaboration Network: 0.529636

Collaboration Network has the largest average clustering coefficient. Considering it has more nodes with higher degrees than the other 2 random graphs, the nodes in the graph are more likely to create tighter clusters and increase the coefficient's value.

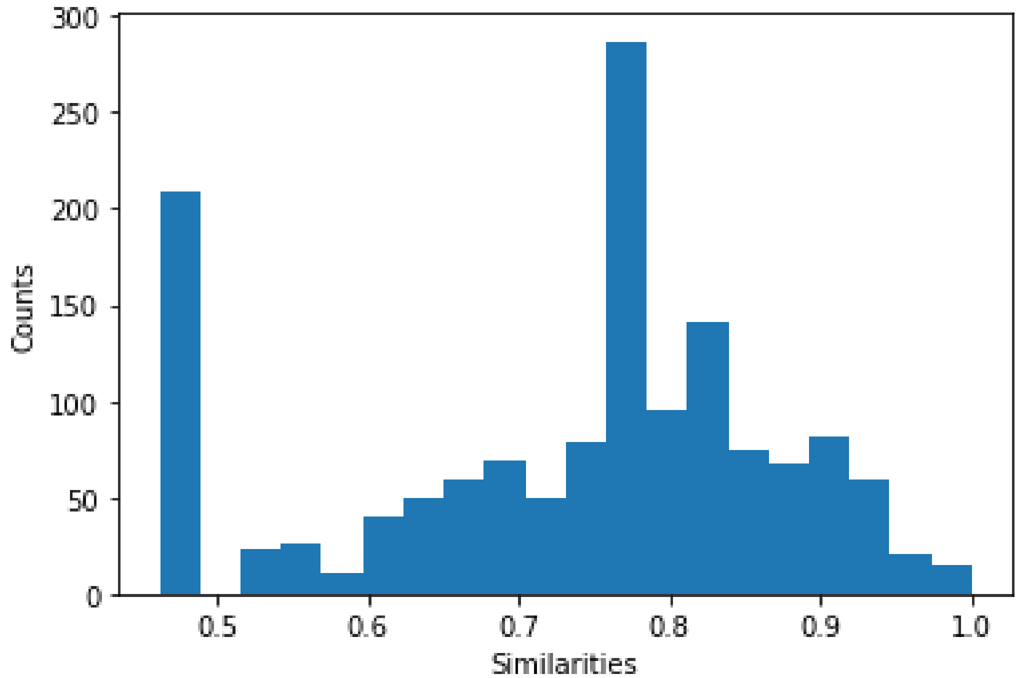
Node 9 feature = [6, 4, 1]

Top 5 similar nodes are: (0.9975670804741605, 415), (0.9905211130872973, 286), (0.9905211130872973, 288), (0.9889960604305609, 1054), (0.9889960604305609, 1336)

Node 9 feature = [6.0, 4.0, 1.0, 2.5, 1.8333333333333333, 4.0, 15.0, 11.0, 24.0, 2.5, 1.8333333333333333, 4.0, 4.319444444444444, 3.027777777777778, 2.388888888888889, 10.166666666666666, 7.333333333333333, 6.5, 15.0, 11.0, 24.0, 25.916666666666664, 18.166666666666668, 14.333333333333332, 61.0, 44.0, 39.0]

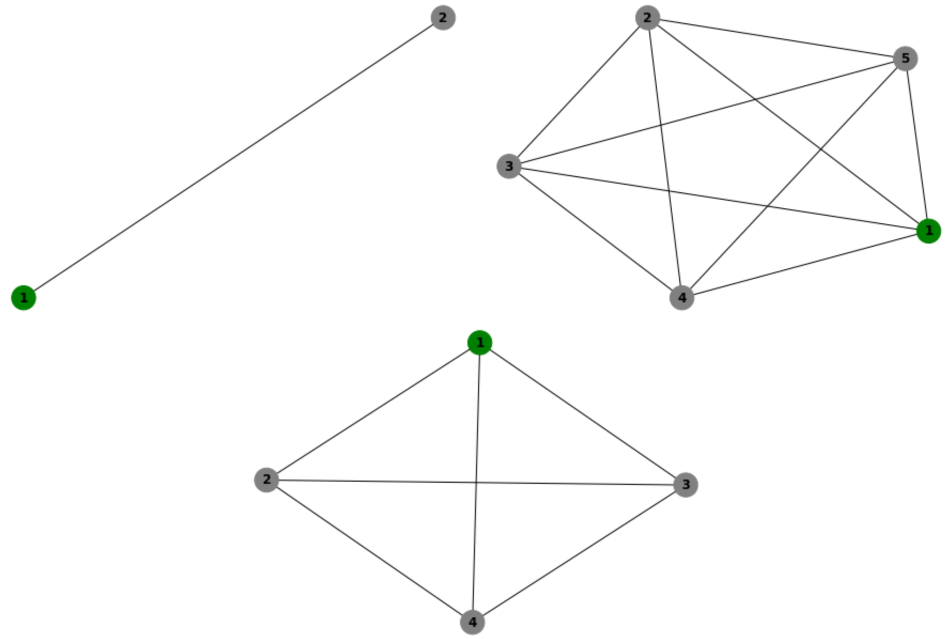
Top 5 similar nodes are: (0.9924963736111307, 973), (0.9901720708825511, 415), (0.986033019147163, 537), (0.9851636059311806, 24), (0.9851636059311806, 25)

There is only 1 common node on the 2 lists. The difference between the two lists stems from the fact that the features used in 2.2 takes into account features of nodes outside of the node's egonet.



There are two very noticeable spikes and one spike that is less visible but still stands out.

The green nodes in the figures below are the centers of the egonets.



The node in the first graph belongs to a group with only binary relationships. The nodes in the last 2 graphs belong to groups of nodes with fully connected egonets but has different number of members.

$$Q = \frac{1}{2m}\left(\sum_{1\leq i,j\leq n} A_{i,j}\delta(c_i,c_j) - \frac{(\sum_{1\leq i,j\leq n} d_i)\delta(c_i,c_j)(\sum_{1\leq i,j\leq n} d_i)\delta(c_i,c_j)}{2m}\right)$$

$$A = \frac{d_i d_j}{2m}\delta_{c_i,c_j\notin(C\cup k)}$$

$$Therefore:$$

$$Q_{before} = \frac{1}{2m}\left(\sum_{in} - \frac{(\sum_{tot})^2}{2m} - \frac{(k_i)^2}{2m} - A\right)$$

$$= \frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2 - A$$

$$Q_{after} = \frac{1}{2m}\left(\sum_{in,k_{i,in}} - \frac{(\sum_{tot} + k_i)^2}{2m} - A\right)$$

$$= \frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2 - A$$

$$Therefore:$$

$$\Delta Q = \left(\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right) - \left(\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right) Q.E.D$$

- 1
- 6
- $4\left(\frac{12}{2*28} - \left(\frac{14}{2*28}\right)^2\right) = 0.60714$
- 2
- 13
- $2\left(\frac{26}{2*28} - \left(\frac{28}{2*28}\right)^2\right) = 0.42857$

- 1
- 6
- $32\left(\frac{12}{2*224} - \left(\frac{14}{2*224}\right)^2\right) = 0.82589$
- 2
- 13
- $16\left(\frac{26}{2*224} - \left(\frac{28}{2*224}\right)^2\right) = 0.86607$

In the above case, the algorithm found a partition that might not be intuitive when looking at individual nodes, but it might make sense when looking at the graph in a larger scale. This is because the algorithm essentially finds a hierarchical of groupings and each pass of its iteration finds a local optima that is meaningful at that level of the hierarchy.

- The resulting nxn matrix resulting from each element of the summation is a sparse matrix with A[i,j] = A[j,i] = -1 and A[i] = A[j] = 1. Thus $\sum_{(i,j)\in E}(e_i - e_j)(e_i - e_j)^T = L$

- Since $x^T(e_i - e_j) = (e_i - e_j)^T x = x_i - x_j$, we have: $x^T Lx = \sum_{(i,j)\in E}(x_i - x_j)^2$ QED

- With $x^T Lx = \sum_{(i,j)\in E}(x_i - x_j)^2$, we can see that if c_i and c_j are in the same set then that element is 0. Thus:
$x^T Lx = a\left(\frac{\sqrt{vol(S)}}{\sqrt{vol(\overline{\mathrm{S}})}} + \frac{\sqrt{vol(\overline{\mathrm{S}})}}{\sqrt{vol(S)}}\right)^2 = a\frac{(vol(S)+vol(\overline{\mathrm{S}}))^2}{vol(S)vol(\overline{\mathrm{S}}}$
$= 2m\left(\frac{a}{vol(S)} + \frac{a}{vol(\overline{\mathrm{S}})}\right) = 2m * NCUT(S)$ QED

- We have $De = d$ which is the degree vector.
Then:
$x^T De = x^t d = \sum_{c_i\in S} d_i \frac{\sqrt{vol(S)}}{\sqrt{(vol(\overline{\mathrm{S}})}} + \sum_{c_i\notin S} d_i \frac{\sqrt{vol(\overline{\mathrm{S}})}}{\sqrt{vol(S)}}$
$= \sqrt{(vol(\overline{\mathrm{S}})}\sqrt{vol(S)} - \sqrt{vol(S)}\sqrt{vol(\overline{\mathrm{S}})} = 0$

- $x^T Dx = vol(S) + vol(\overline{\mathrm{S}}) = 2m$

Set $z = xD^{1/2}$. We have an equivalent problem:

$$min_{z \in R^n} \frac{z^T \hat{L} z}{z^T z}$$
$$s.t : z^T D^{1/2} e = 0, z^T z = 2m$$

We can write z in basis of eigenvectors $w_1, w_2, .., w_n$ of with corresponding eigenvalues λ_i.

Therefore $z^T \hat{L} z = \sum_i \lambda_i \alpha_i^2$

We also have $D^{1/2}e$ as the smallest eigenvector of of . So the restrain $z^T D^{1/2} e = 0$ holds when $\lambda_1 = 0$ and $\sum_i \alpha_i^2 = 2m$.

So to minimize the problem, we need to set $\lambda_2 = \sqrt{2m}$ and the rest to 0. Which means $z = \sqrt{2m} w_2$, so $x = \sqrt{2m} D^{-1/2} w_2$. We can set c_i in S if $x_i > 0$ to satisfy the original problem.

$$
\begin{aligned}
Q(y) &= \frac{1}{2m}\left(\sum_{i,j} A_{i,j} I_{y_i=y_j} - \sum_{i,j} \frac{d_i d_j}{2m} I_{y_i=y_j}\right) \\
&= \frac{1}{2m}\left(2m - 2cut(S) - \frac{vol(S)^2 + vol(\overline{\mathrm{S}})^2}{2m}\right) \\
&= \frac{1}{2m}\left(-2cut(S) + \frac{(vol(S) + vol(\overline{\mathrm{S}}))^2 - vol(S)^2 - vol(\overline{\mathrm{S}})^2}{2m}\right) \\
&= \frac{1}{2m}\left(-2cut(S) + \frac{1}{m}(vol(S) + vol(\overline{\mathrm{S}}))\right) QED
\end{aligned}
$$

Information sheet
CS224W: Machine Learning with Graphs

Assignment Submission Fill in and include this information sheet with each of your assignments. This page should be the last page of your submission. Assignments are due at 11:59pm and are always due on a Thursday. All students (SCPD and non-SCPD) must submit their homework via GradeScope (http://www.gradescope.com). Students can typeset or scan their homework. Make sure that you answer each (sub-)question on a separate page. That is, one answer per page regardless of the answer length. Students also need to upload their code on Gradescope. Put all the code for a single question into a single file and upload it.

Late Homework Policy Each student will have a total of *two* late periods. *Homework are due on Thursdays at 11:59pm PT and one late period expires on the following Monday at 11:59pm PT*. Only one late period may be used for an assignment. Any homework received after 11:59pm PT on the Monday following the homework due date will receive no credit. Once these late periods are exhausted, any assignments turned in late will receive no credit.

Honor Code We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently, i.e., each student must understand the solution well enough in order to reconstruct it by him/herself. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web (GitHub/Google/previous year's solutions etc.) is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code very seriously and expect students to do the same.

Your name: Truong Pham ______________________________

Email: ______________________________ **SUID:** ______________________

Discussion Group: ____

I acknowledge and accept the Honor Code.

(Signed) ____